Advanced Information Systems Technologies (AIST)

Machine Learning in NASA's Earth Science Division

Michael Little

February 21, 2018

ESTO
Earth Science Technology Office

Outline of Talk

- Hope, Tragedy, and Myths
- What’s it Going to Take
- Analytic Center as a Framework
- AIST Managed Cloud Environment (AMCE)
- Arc-GIS Online (AGOL)
- Lessons Learned
- Backup

Hope,

- A more robust understanding of a natural phenomenon or physical process
  - Relying on conclusions based on 3,000 examples, not 3
  - Digest massive volumes of observational data
  - High volume output from high-resolution ES models
  - Conventional techniques drown ES researchers
- Prediction of future state for complex systems from observational data without a thorough understanding of complex or poorly understood underlying physics
- Coordination of elements of a sensor web to target transient and transitional phenomena
  - Leverage emergence of SmallSats
  - Autonomy
  - Detection of interesting features and re-tasking to observe them
- Fast comparative OSSEs
- Assimilation models which do better at washing out the experimental error while preserving features and physical anomalies

Tragedy,

- Large investments by DoD and IC in AI and Machine Learning have, largely, not been leveraged by NASA Earth Science
- Lack of credibility in ML by Earth Science researchers
  - Why does ML generate bogus conclusions?
    - Failure to articulate assumptions & constraints and exceeding them
    - Overtraining
    - Drift over time
    - Failure to recognize when conditions change
  - Earth Scientists have trouble understanding the technology
    - Lack of demonstration of value to motivate them
    - Lack of training in the tools and algorithms
- Partnerships among Machine Learning and Earth Science communities are slow to take root
  - Communications
  - Opportunity
  - Trust
- Data grooming takes huge amounts of time

Myths

- Do a problem, publish for others to use, move on to next.
  - Reality: There is no fire and forget in Earth Science
- Open source is always better than commercial software
  - GIS is a good counter-example
  - Communicating with other Agencies requires arc-GIS capability
- Data must be supplied in same form as collected
  - Reality: Most investigators transform data once they get it
    - Often fail to verify measurements, without introducing errors
  - Must be able to trace back to authoritative data
- Cannot re-use (trust) anyone else’s work
  - Community acceptance, peer review could create common capabilities
    - Journal publications want to help with this
    - Apache wants to help with this
  - Fix the data once and others could use it if they believe the fix

What's it Going to Take?

- **Community Acceptance** in NASA Earth Science (ES)
  - NASA Earth Scientists must understand the processes & algorithms
    - "In the field of observation, fortune favors the prepared mind." – Pasteur
    - Collaboration among Earth Scientists, Computer Scientists and Technologists
  - Demonstration of value to the science communities
    - New Science
    - Faster Science
    - Results backed up by confirmation with legacy techniques
    - Uncertainty Quantification
  - Engagement with NASA ES by experienced ML practitioners
- **Easy to use** tools and credible results
  - To Earth Scientists
  - Which make available accepted processes and algorithms
- Improved **Data Usability**
  - Rapid use of the numeric results without significant grooming

What is an Analytic Center

- An environment for conducting a Science investigation
  - Enables the confluence of resources for that investigation
  - Tailored to the individual study area
- Harmonizes data, tools and computational resources to permit the research community to focus on the investigation
  - Reduce the data preparation time to something tolerable
  - Catalog of optional resources (think HomeDepot shopping)
  - Semantic-enabled catalog of resources (think Yelp) with help
  - Relevant publications
  - Provide established training data sets of varying resolution
  - Provide effective project confidentiality, integrity and availability
  - Single sign-on and unified financial tracking

Analytic Center as a Framework

User

- Project Definition
- Plan for Investigation

Data

- Catalog of repositories
- NASA DAAC
- Other US Govt
- Non-US
- Local or non-public

Project Work Environment

Tools

- Discovery
- Catalog of assets
- Work Management
- Data Interfaces
- Analytic Tools
- Modeling
- Collaboration
- Visualization
- Sharing/ Publication
- Local/custom tools
- Help

Computational Infrastructure

- Computing
  - Capacity
  - Capability
- Storage
- Communications

Storage

- Data Containers
- Thematic model
- Metadata/Ontology
- Resulting Products
- Published data
- Provenance

Computing

- Local systems
- High End Computing
- Cloud Computing Capability
- Quantum Computing
- Neuromorphic Computing

ESTO
Earth Science Technology Office

OceanWorks as an Example Analytic Center

Physical Oceanographers

- Project Definition
- Plan for Investigation

Data

- Earthdata CMR
- nonCMR DAAC
- PI Generated
  - ECCO
  - Altimetry
- EDGE
  - Earthdata CMR
  - nonCMR DAAC
  - PI data ingest
  - GRACE (CSR)
  - ECCO (MIT)
  - Altimetry (OSU)
- DOMS
  - ICOADS
  - SAMOS
  - SPURS
  - ASCAT Winds
  - Aquarius
  - MODIS SST
  - MUR-SST

Project Work
Environment
-OceanWorks-

Tools

- **EDGE and MUDROD: Metadata search/Discovery**
- **Services**
  - Area Averaged Time series
  - Time averaged map
  - Correlation Map
  - Daily Differences
  - Matchup (single satellite-multiple in situ)
- **Workflow**:
  - Jupyter Notebook
  - Zeppelin Notebook
- **Visualization**
  - OnEarth
  - CMC (GIS)
- **Collaboration**
  - Confluence, JIRA, GIT, Bamboo
  - Apache wiki
  - Smartsheet and Google Office
  - Slack
- Expert help - MUDROD

Computational Infrastructure

Storage

- Nexus
- Apache Solr
- Amazon S3

Computing

- Local systems
- JPL AWS
- AMCE Cloud Computing
- JPL on premises Cloud
- NGAP

POC: Thomas.huang@nasa.gov

Analytic Center Characteristics

- Little to NO overhead when not being used
  - Persistent sources of funding to support infrastructure for science are hard to find
- Seamless integration of new/custom tools
- Comprehensive catalog of tools and known data (from many repositories)
  - Clear applicability (or not)
  - Expert system as an operator aid
- Help in using them
  - YouTube, moodle

Theory of Data Systems (ToDS)

- ToDS is SAMSI collaboration on how should data systems support analysis
  - NASA, JPL, NOAA, NSF collaborating to support ToDS
  - NASA Point of Contact: Amy Braverman at JPL
- Primary Focus is on **Uncertainty Quantification** in multi-source data analysis
- Most Science Investigations need data from **MULTIPLE SOURCES**
  - Each data supplier focuses on their own data
  - Different geolocation reference point in each data product
  - Different flags for no-observation
  - Quality flags instead of error but often no uncertainty quantification
  - Different grid references
  - Different statements/estimates of uncertainty and error
- NASA Mission design focuses on stewardship and Cal/Val
  - Format is unique to the instrument science team and the data product
- Large volumes of data from disparate sources require modern analytic techniques, tools and computing capacity
  - Regression techniques lack techniques to quantify uncertainty

Computing Resources

- Analysis of Earth Science phenomena requires large scale computing
  - Analysis of observations of many events (examples of phenomena or processes) requires large quantities of computing
  - Modeling often requires highly interconnected nodes to handle non-linear phenomena
  - Visualization requires appropriate numeric files to make meaning obvious
- Computing choices available
  - NASA High-end Computing Centers (NAS and NCCS)
  - Other Government supercomputing sites in conjunction with colleagues
  - Commercial Cloud Computing
    - AWS
    - Google
    - Microsoft
    - Suppliers of specialized cloud computing
- Use of computing requires funding and an Authorization to Operate (ATO)
  - Most investigations are in the LOW category
    - Limited NASA need for Moderate category in Science investigations
  - Many require collaboration with foreign nationals, which is hard to do inside the NASA address space

AMCE Management Considerations

- AIST demonstrated that Scientific Investigations can be conducted by NASA in a public cloud computing environment
  - Must support the way in which NASA ESD scientific investigations are conducted
- Analogous to developing an apartment building
  - Provide maximum flexibility for residents without impacting other residents
  - Not all projects belong in a multi-tenant facility
  - Some projects need to be in a single family dwelling
- Fundamental Principles
  - Provide Project PI with maximum flexibility in determining what to use in AWS
  - Ensure that projects are protected from each other
  - Help the PI ensure that his projects do not over run their budget
    - Automate financial overrun notices: alert, warning and shutdown
  - Monitor compliance with essential Computer Security restrictions
    - Automate notification of problems
- AIST assumes NASA Earth Science Data will be supplied via EOS-DIS efforts
  - AMCE provides a complementary service for PI's to buy computing near the data
  - AMCE is not intended to provide a repository for data
- NCCS is in process of operationalizing AMCE
  - Eventual expansion into Google and Microsoft Azure environments

PI's Receive Benefits of Cloud Computing

- Full Range of NASA Investigation Teams on a single computing environment
  - University, JPL, NASA Center, Other Government Agencies and Industry
  - Supercomputing requirements filled through normal HEC processes
- Startup Leadtime can be as little as 5 days
- Multi-party collaboration in neutral environment
  - NASA and non-NASA individuals and organizations
  - Exists outside NASA address space so no opportunity to contaminate NASA systems
- Use funding to buy time on AMCE instead of authorizing purchase of hardware
  - AIST directly funds transition and learning curve
  - Research Project funds consumption of AWS services after transition with 10% overhead
  - History indicates this cost is typically far lower (i.e., 10% or less) of investment cost
  - Other partners can buy their own AWS time and yet integrate into the computing environment
- Codes developed in AMCE are more easily integrated into NASA computing systems
- Similar needs elsewhere in Science Mission Directorate (SMD) Research divisions

Experiments with an ATO for arc-GIS Online (AGOL)

- AGOL is a cloud-based version of arc-GIS online allowing for collaboration with non-NASA data, users and colleagues
- Accepted challenge due to widespread requests for accounts on Esri Enterprise License Agreement (ELA)
  - Users were unable to obtain an ATO for their individual projects
  - FEDRAMP Process was stalled (although moving at this time)
- Opportunity to experiment with supplying data from arc-GIS Enterprise Servers to AGOL
- ATO Letter is expected to be ready for signature within two weeks (3/1/18)
  - ATO for Software-as-a-Service is unfamiliar ground at NASA
  - Approval expected at LOW Computer Security rating
    - ONLY Public Data

AIST Strategy

- Create an environment for conducting scientific investigations with data, tools and computing resources
  - Machine Learning workshop in April - ESIP
- Fund collaborations among scientific investigators, data experts and analytics experts
  - AIST-18 will fund experiments in the use of data analytics, not just technology development
  - Multi-disciplinary teams with the goal of enabling science
  - Draft solicitation in April, 2018 for community comment

AC Considerations in CA

- Focus must be on enabling individual scientific investigations
  - Maximum flexibility must be afforded to the PI
  - Accelerate the process of identifying, acquiring and cleaning up the data for their use
  - Enable access to both NASA and non-NASA data and tools
- Framework must make it easy to add new tools and capabilities without complex integration
  - Most investigators prefer the tools they are comfortable with and understand
  - Must be able to integrate contributions from others
- Framework must have very low overhead cost
  - No cost to keep it around when not being used
- Role of the DAACs in Analytic Centers
  - Provide effective, fast interfaces to the data products they hold
    - Allow scientists to extract the data of interest from archive
  - Provide assistance in the analysis and interpretation of their data products
  - Curation of well-documented Training Sets for Supervised Machine Learning
  - Support the science communities for their themes
    - Aid in the development of analytic tools related to their thematic responsibilities

Backup

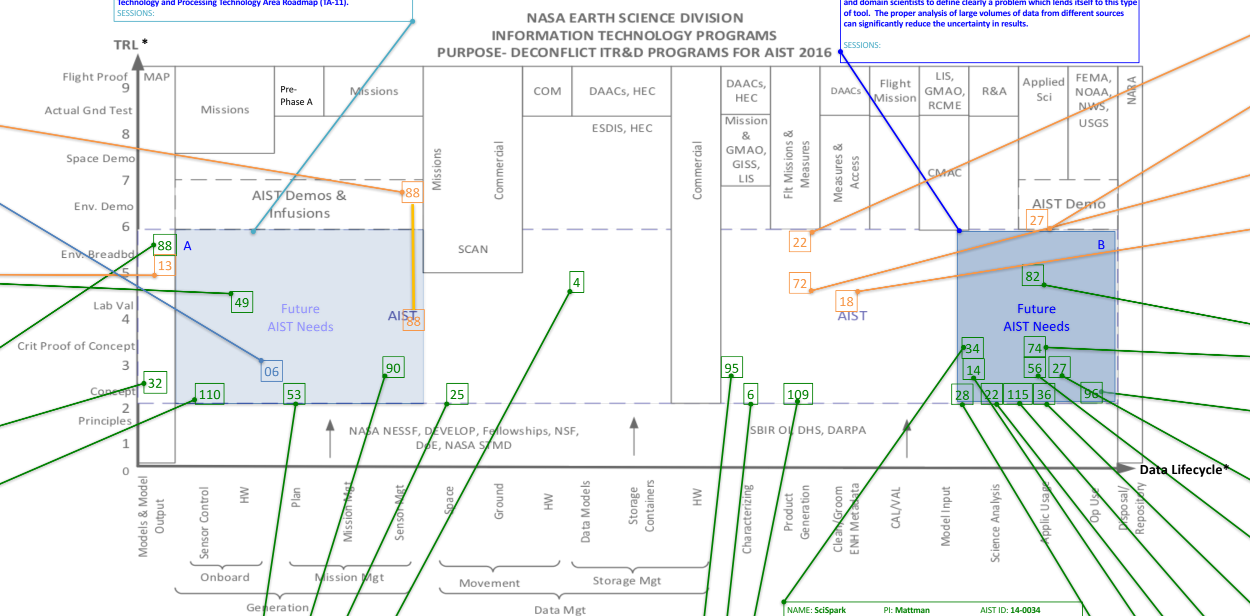
- AIST
  - Responsibilities
  - AIST Projects with Machine Learning Components
- A Sample of Indian Projects using Machine Learning
- Machine Learning in China
- Machine Learning in Europe
- Useful Tools and Algorithms
- Analytic Center Needs

AIST Responsibilities

AIST Projects Plotted Against their TRL and Relevant Phase of the Data Life Cycle

NAME: Flight Control PI: Future AIST ID: A
TRL: 2 to 6 Phase: Generate-Sensor/Spacecraft Control
TITLE: Requirements for Sensor/Spacecraft Control IT Capabilities
There is a growing demand for measurements of transient phenomena that requires real-time aiming/pointing/staring of instruments or coordination among spacecraft with only limited human supervision. Planning as well as operational control are needed. These require the ability to automate operations using instrument output without the latency restrictions imposed by earthbound processing. Further information can be obtained in NASA's Technology Roadmaps, particularly the Modeling, Simulation, Information Technology and Processing Technology Area Roadmap (TA-11).
SESSIONS:

NAME: Data Analysis PI: Future AIST ID: B
TRL: 2 to 6 Phase: Exploit- Data Analysis
TITLE: Requirements for IT Data Analysis Capabilities
The growth in volume of NASA Earth Science Data and Model Output necessitate the use of automated analysis tools. The value of machine learning and artificial intelligence in sifting through large volumes of dissimilarly formatted data has established credibility in particle physics and financial markets. Limited initial experiments in earth science have demonstrated the ability to improve the quality and statistical basis of scientific discovery in this domain, also. Successful projects involve collaboration between data scientists and domain scientists to define clearly a problem which lends itself to this type of tool. The proper analysis of large volumes of data from different sources can significantly reduce the uncertainty in results.
SESSIONS:



NAME: SoilScape PI: Moghaddam AIST ID: 11-0088
TRL: 7 Phase: Exploit-Secondary Products
TITLE: Soil moisture Sensing Controller & oPtimal Estimator (SoilSCAPE)
SESSIONS: N/A

NAME: HPSC PI: Doyle AIST ID: 16-QRS-0006
TRL: 2 Phase: Generate-Flight Computing
TITLE: High Performance Spacefight Computing (HPSC) Capability
SESSIONS: None

NAME: LIS PI: Peters-Lidard AIST ID: 11-0013
TRL: 6 Phase: Exploit-Secondary Products
TITLE: A mission simulation and evaluation platform for terrestrial hydrology using the NASA Land Information System (LIS)
SESSIONS: H41L-02

NAME: SpaceCubeX PI: French AIST ID: 14-0049
TRL: 3 to 5 Phase: DM-Characterize
TITLE: SpaceCubeX: A Hybrid Multi-core CPU/FPGA/DSP Flight Architecture for Next Generation Earth Science Missions
SESSIONS: IN41C-1707

NAME: Large ensemble PI: Clark AIST ID: 14-0088
TRL: 3 to 5 Phase: Generate – Model Management
TITLE: Development of Computational Infrastructure to Support Hyper-resolution Large-ensemble Hydrology Simulations from Local-to-Continental Scales)
SESSIONS: H23E-1620, H33E-1655, H13R-03

NAME: CMDA PI: Lee AIST ID: 14-0032
TRL: 3 to 5 Phase: Generate: Model Management
TITLE: Climate Model Diagnostic Analyzer (CMDA)
SESSIONS: IN31A-1753, IN31A-1754, IN31A-1753

NAME: MPC PI: Lieber AIST ID: 14-0110
TRL: 2 to 4 Phase: Generation- Sensor Control
TITLE: Model Predictive Control Architecture for Optimizing Earth Science Data Collection
SESSIONS:

NAME: iISCE PI: Rosen AIST ID: 11-0022
TRL:6 Phase: Exploit-Secondary Products
TITLE: iISCE: InSAR Scientific Computing Environment on the Cloud. Additionally SAR SCP Foundry
SESSIONS: IN43B-1734

NAME: Hurricanes PI: Hristova-Veleva AIST ID: 11-0027
TRL:6 Phase: Exploit-Secondary Products
TITLE: Fusion of Hurricane Models and Observations: Developing the Technology to Improve the Forecasts
SESSIONS: IN51C-05, IN51B-1811

NAME: PI: Bock AIST ID: 11-0072
TRL: 5 Phase: Exploit-Secondary Products
TITLE: Next-Generation Real-time Geodetic Station Sensor Web for Natural Hazards Research and Applications
SESSIONS:

NAME: Data Fusion PI: Braverman AIST ID: 11-0018
TRL:3 Phase: Exploit-Secondary Products
TITLE: Multivariate Data Fusion and Uncertainty Quantification for Remote Sensing
SESSIONS: A41-0192

NAME: MUDROD PI: Yang AIST ID: 14-0082
TRL:5 to 7 Phase: Exploit-Data Analysis
TITLE: Mining and Utilizing Dataset Relevancy from Oceanographic Dataset (MUDROD) Metadata, Usage Metrics, and User Feedback to Improve Data Discovery and Access
SESSIONS: IN51B-1809, IN43B-1735, IN31C-1774, IN51B-1812, IN41D-06, PA53A-2237, IN43B-1726, IN31C-1772

NAME: Flooding PI: Brakenridge AIST ID: 14-0074
TRL: 3 to 5 Phase: Exploit: Application Use
TITLE: Global Flood Risk From Advanced Modeling and Remote Sensing in Collaboration With Google Earth Engine
SESSIONS: IN31D-05, IN13D-1968

NAME: Quantum PI: Halem AIST ID: 14-0096
TRL: 2 to 4 Phase: Generate Compute Environment
TITLE: Feasibility Studies of Quantum Enabled Annealing Algorithms for Estimating Terrestrial Carbon Fluxes from OCO-2 and the LIS Model
SESSIONS: IN31A-1758, IN51B-1813

NAME: Pattern-based GIS PI: Stepinski AIST ID: 14-0027
TRL: 3 to 5 Phase: Exploit-Data Analysis
TITLE: Pattern-based GIS for Understanding Content of very large Earth Science datasets
SESSIONS:

NAME: DERECHOS PI: Kuo AIST ID: 14-0056
TRL: 3 to 5 Phase: Exploit-Data Analysis
TITLE: DERECHOS: Data Environment for Rapid Exploration and Characterization of Organized Systems
SESSIONS: IN41C-1707∂

NAME: Discovery PI: Pankratius AIST ID: 14-0036
TRL: 2 to 5 Phase: Exploit-Data Analysis
TITLE: Computer-Aided Discovery of Earth Surface Deformation Phenomena
SESSIONS: IN41C-1707∂

NAME: Climate Impact PI: Das AIST ID: 14-0115
TRL: 4 to 5 Phase: Exploit-Data Analysis
TITLE: Uncovering Effects of Climate Variables on Global Vegetation
SESSIONS: IN41C-1707∂

NAME: CRM PI: Tao AIST ID: 14-00022
TRL: 2 to 5 Phase: Exploit-Storage
TITLE: Empowering Data Management, Diagnosis, and Visualization of Cloud-Resolving Models by Cloud Library upon Spark and Hadoop
SESSIONS: IN51A-1794

NAME: NAIADS PI: Lukashin AIST ID: 14-0014
TRL: 3 to 6 Phase: Exploit-Data Analysis
TITLE: NASA Information And Data System (NAIADS) for Earth Science Data Fusion and Analytics
SESSIONS:

NAME: AMIGHO PI: Larson AIST ID: 14-0004
TRL: 4 to 5 Phase: DM-Stewardship
TITLE: Automated Metadata Ingest for GNSS Hydrology within OODT
SESSIONS: IN41C-1707∂

NAME: SciSpark PI: Mattman AIST ID: 14-0034
TRL: 3 to 5 Phase: DM-Movement
TITLE: SciSpark: Highly Interactive and Scalable Model Evaluation and Climate Metrics for Scientific Data and Analysis
SESSIONS: IN53C-01

NAME: Mission Design PI: Lemoigne AIST ID: 14-0053
TRL: 2 to 4 Phase: Generation: Mission Planning
TITLE: Tradespace Analysis Tool for Designing Earth Science Distributed Missions
SESSIONS:

NAME: Multi-channel PI: Ivancic AIST ID: 14-0025
TRL: 2 to 6 Phase: DM-Movement from Space
TITLE: Multi-channel Combining for Airborne Flight Research using Standard Protocols
SESSIONS:

NAME:SAR Analytics PI: Hua AIST ID: 14-0109
TRL: 3 to 4 Phase: Exploit-Product Generation
TITLE: Agile Big Data Analytics of High-Volume Geodetic Data Products for Improving Science and Hazard Response
SESSIONS:

NAME: UAV PI: Campbell AIST ID: 14-0090
TRL: 4 to 5 Phase: Generation: Sensor Management
TITLE: Next Generation UAV Based Spectral Systems for Environmental Monitoring
SESSIONS: IN41C-1707∂

NAME: ODYSSEUS PI: Gleason AIST ID: 14-0095
TRL: 3 to 5 Phase: DM-Characterize
TITLE: Ontology-based Metadata Portal for Unified Semantics (OlyMPUS)
SESSIONS: IN41C-1707

NAME: Dark Data PI: Ramachandran AIST ID: 14-0006
TRL: 2 to 3 Phase: DM-Characterize
TITLE: Illuminating the Darkness: Exploiting Untapped Data and Information Resources in Earth Science
SESSIONS: IN23E-07

NAME:OceaneXtremes PI: Huang AIST ID: 14-0028
TRL: 2 to 4 Phase: Exploit: Data Analysis
TITLE: Oceanographic Data-in-Analysis Portal
SESSIONS: IN51B-1910, IN23D-1755, In51B-1810, IN31C-1771, In31C-1783, IN51B-1812, IN21D-1709

Legend

49 — AIST Project Sequence Number (Green=AIST14) (Orange=AIST11) (Blue=Future Needs)

Applied Sci	NASA's Earth Science Division (ESD) Applied Sciences Program
Breadbd	Breadboard prototype
CAL/VAL	Calibration/Validation of data products and instruments
CMAC	NASA's Computational Modeling & Algorithm & Cyberinfrastructure Program
COM	Commercial
Crit	Critical
DAAC	NASA's Digital Active Archive Centers
ESDIS	NASA's Earth Science Data and Information System Project
Env	Expected operating environment for the System
FEMA	US Federal Emergency Management Agency
FLT	Flight
GISS	GISS
GMAO	Goddard Modeling and Assimilation Office
Gnd	Ground
HEC	NASA's High End Computing Program
HW	Hardware
LIS	NASA's Land Information System at Goddard Spaceflight Center
Lab Val	Validation in a laboratory environment
MAP	NASA's Modeling and Analysis Program
MEASURES	M
Mgt	Management
NARA	US National Archives and Records Administration
NOAA	US National Oceanographic and Atmospheric Administration
NWS	US National Weather Service
Op Use	Operational Use
Pre-OA	Pre-Phase A Program
R&A	NASA's Earth Science Division Research and Analysis Programs
RCME	Regional Climate Modeling environment
SCAN	NASA's Space Communications and Navigation Program
TRL	Technology Readiness Level
USGS	US Geological Survey

12/21/15

*Notes and References:
1. **TRL** is a commonly used measure of the maturity of a technology. The criteria for successive levels is described at http:://esto.nasa.gov/files/trl_definitions.pdf

2. Data Life Cycle Phases are described at 2015 IEEE Big Data Conference in Santa Clara, CA. More information available at **http://ieee-bigdata-earthscience.jpl.nasa.gov/**

ESTO Earth Science Technology Office

A Possible Framework for Adoption

- Focus on needs of a particular Science community
  - Data, models and tools they are comfortable with
  - Computational resources adequate to perform their investigation
  - Concept of Operations must support their work
- NASA’s Earth Science needs **much** more exposure to Machine Learning
  - Researchers and Managers need to develop confidence in the tools
  - Outreach to the Science community
- Must be an architecture with modularity and interoperability
- Must support security
- Increase access to non-NASA data sources as well as NASA
- AIST has been experimenting with basic concepts as well as components

AIST Experiments with Analytic Centers

- Land Use and Land Change
  - NEX (NASA Ames) Rama Neimani
- Physical Oceanography
  - OceanWorks (JPL) Thomas Huang
- Tropical Cyclones and Hurricanes
  - TCIS (JPL) Svetla Hristova-Veleva
- Climate
  - Climate Model Diagnostic Analyzer (CMDA), Seungwon Lee (JPL)
  - Climate Workbench (UAH/MSFC) Manil Maskey, Chris Lynnes
- Communities being discussed for further experiments
  - Biodiversity
  - Hydrology
  - Cryosphere

AIST-14 Tasks-Machine Learning

PI Name	Org	Prop #	Project Title	ESTO PM	End Date	Machine Learning Elements
Kamalika Das	UC Santa Cruz	115	Uncovering Effects of Climate Variables on Global Vegetation	Oza	5/31/17	Symbolic regression
Milton Halem	*UMBC*	*96*	Computational Technologies: Feasibility Studies of Quantum Enabled Annealing Algorithms for Estimating Terrestrial Carbon Fluxes from OCO-2 and the LIS Model	Cole	5/31/17	RBM, CNN, data assimilation
Hook Hua	JPL	109	Agile Big Data Analytics of High-Volume Geodetic Data Products for Improving Science and Hazard Response	Norton	5/31/17	Fault recognition and Science processing redirection
Thomas Huang	JPL	28	OceanXtremes: Oceanographic Data-Intensive Anomaly Detection and Analysis Portal	Norton	8/29/17	NEXUS as a data delivery tool - oceanography
Kristine Larson	U of CO	4	AMIGHO: Automated Metadata Ingest for GNSS Hydrology within OODT	Hines	5/31/17	Sensor characterization
Victor Pankratius	MIT	36	Computer-Aided Discovery of Earth Surface Deformation Phenomena	Little	5/31/17	Computer aided discovery
Chaowei Yang	GMU	82	Mining and Utilizing Dataset Relevancy from Oceanographic Dataset (MUDROD) Metadata, Usage Metrics, and User Feedback to Improve Data Discovery and Access	Cole	5/31/17	Natural language processing, CNN, SVM, deep learning
Tomasz Stepinski	Cincinnati	27	Pattern-based GIS for Understanding Content of very large Earth Science datasets	Quam	6/31/17	Classification and similarity
Jonathan Gleason	LaRC	95	Ontology-based Metadata Portal for Unified Semantics (OlyMPUS)	Oza	1/31/17	Precision ontology foundation
Constantine Lukashin	LaRC	14	NASA Information And Data System (NAIADS) for Earth Science Data Fusion and Analytics	Murray	1/31/17	Scientific enterprise service bus
Aashish Chaudhary	Kitware	65	Visualization Pipelines for big-data on the NASA Earth Exchange (NEX) Prototyping Agile Production, Analytics	Hines	3/31/17	Workflow
Martyn Clark	UCAR	88	Development of Computational Infrastructure to Support Hyper-resolution Large-ensemble Hydrology Simulations from Local-to-Continental Scales	Hines	4/30/17	Assimilation and ensembles
Kwo-Sen Kuo	Bayesics	56	DERECHOS: Data Environment for Rapid Exploration and Characterization of Organized Systems	Little	4/30/17	Foundation for data delivery
Seungwon Lee	JPL	32	Climate Model Diagnostic Analyzer	Norton	4/30/17	
Christian Mattman	JPL	34	SciSpark: Highly Interactive and Scalable Model Evaluation and Climate Metrics for Scientific Data and Analysis	Norton	4/30/17	Foundation for data delivery

AIST QRS Tasks-Machine Learning

PI Name	Org	Prop #	Project Title	ESTO PM	End Date	Machine Learning Elements
Dan Duffy	GSFC	AIST-QRS-16-0002	Demo of VR Technology with live data(Phase 2) & GIS (Phase 3)	Cole	12/31/16	VR
Sreeja Nag	ARC	AIST-QRS-16-0003	Scheduling Satellite Pointing within Constellations	Hines	12/31/16	
Ved Chirayeth	ARC	AIST-QRS-16-0004	MiDAR-fused Supervised Machine Learning (SML)	Hines	1/31/17	Use of high resolution training sets to improve global scale moderate resolution data
Chris Mattman	JPL	AIST-QRS-16-0007	Deep Web Search Analytics	Norton	10/31/16	Text analytics
John Readey	HDF Group	ACCESS15 -0031	Object Store-based Data Service for Earth System Science	Hines	5/31/17	Foundation for data storage to improve access
Yehuda Bock	Scripps	AIST-QRS-16-0010	Latency test of realtime warning systems in AMCE Cloud Computing	Quam	9/30/17	Feature Detection in near real time
Amy Braverman	JPL	AIST-QRS-16-0005	Probablistic Climate Model Evaluation	Norton	9/20/16	What do you mean by similar?
Chris Lynnes	GSFC	AIST-QRS-16-0001	Experiment with Data Containers in ESDIS Context	Little	12/20/16	Evaluation of alternative storage models for data to enable analysis

AIST-16 Tasks-Machine Learning

PI Name	Org	Prop #	Project Title	ESTO PM	Start Date	End Date	Machine Learning Elements
Victor Pankratius	***MIT***	***48***	Computer Aided Discovery and Algorithmic Synthesis for Spatio-Temporal Phenomena in InSAR	Oza	8/1/17	**7/30/19**	Computer Aided Discovery
Jacqueline LeMoigne-Stewart	**GSFC**	**107**	TAT-C ML	Quam	8/1/17	**7/30/19**	Design space iterator
Branko Kosovic	**NCAR**	**79**	Fuel moisture content for improved fire prediction	Hines	6/1/15	**5/31/17**	Data assimilation
Andrew Michaelis	**ARC**	**137**	Framework for Mining and Analysis of Petabyte-size Time-series on the NASA Earth Exchange (NEX)	Oza	8/1/15	**7/31/17**	Time series analysis
Barton Forman	**UMD**	**24**	A mission planning tool for next generation remote sensing of snow	Quam	9/1/17	**8/30/19**	OSSE and assimilation
Ved Chirayath	***ARC***	***48***	NeMO-Net - The Neural Multi-Modal Observation & Training Network for Global Coral Reef Assessment	Oza	9/1/17	**8/31/19**	High resolution training sets improve moderate resolution imagery
Dara Entekhabi	**GSFC**	**49**	Autonomous Moisture Continuum Sensing Network	Hines	9/1/17	**8/31/19**	
Milton Halem	**UMBC**	**91**	An Assessment of Hybrid Quantum Annealing Approaches for Inferring and Assimilating Satellite Surface Flux Data into Global Land Surface Models.	Little	6/1/15	**8/31/19**	Assimilation, registration
Jeffrey Morisette	**USGS**	**124**	Advanced Phenological Information System	Hines	9/1/17	**8/31/19**	
Walter Jetz	**Yale**	**92**	Software workflows for remote sensing-based biodiversity change monitoring	Cole	9/15/17	**9/14/19**	
Jonathon Hobbs	**JPL**	**30**	Simulation-based Uncertainty Quantification	Norton	10/1/17	**9/30/19**	Statistics
Martyn Clark	**NCAR**	**81**	Climate risks in the water sector: Advancing the readiness of emerging technologies in climate downscaling and hydrologic modeling	Hines	10/1/17	**9/30/19**	

Other AIST Activities Relevant to Machine Learning

- **AIST Virtual Reality Study**
  - Report Completed and undergoing revision
  - To be published by September 30, 2017
  - PoC: Shayna Skolnik (NAVTECA) and Emily Law (JPL)

- **AIST Block-chain Study**
  - Report completed 8/10/17 and undergoing revision
  - To be published by September 30, 2017
  - Interest in results by NRO, NSA, HEOMD (JSC, GRC), SCaN
  - PoC: Chris Dwan

- **Community Driven Workflow Recommendations**
  - CMAC11-0056 (funded by AIST) and AIST14-0032 (Seungwon Lee)
  - PoC: Jia Zhang, Carnegie Mellon University

- **AIST Machine Learning Workshop**
  - Fantastic Problems and Where to Find Them
  - Early CY18
  - Current PoC: Dan Crichton, JPL

- AIST-18 Solicitation (ROSES18)

Non-Government US Technology Investments

- https://www.hpcwire.com/2017/08/17/google-releases-deeplearn-js-democratize-machine-learning/
- https://www.blog.google/topics/machine-learning/pair-people-ai-research-initiative/

Candidate Useful ML Tools

- Implementations
  - TensorFlow
  - TensorFire (ML in the browser using jess)
  - arc-GIS server (Esri)
  - MathLab (Mathworks)
- Algorithms
  - Logistic regression
  - Symbolic regression
  - Random forest
  - Convolutional Neural Networks (CNN)
  - Deep Neural Networks (DNN)
  - Case-based reasoning
  - Restricted Boltzman Machines

Tropical Cycle Information System (TCIS)

Tropical Cyclone Research

- Project Definition
- Plan for Investigation

Data

- Integrated web portal
- Hurricane List & Tracks
- HS3 Field Campaign
- SMAP
- SST
- TPW
- WIND
- ECMWF output
- HAMSR
- GOES-East
- GOES-West
- Microwave Rain Signature

Project Work Environment -JPL-

Tools

- Integrated through web portal
- Re-use requires TCIS Team
- NEOS3 instrument simulator
- Statistical evaluation
- Storm Structure
- Environment vertical structure
- Visualization
- Models

Computational Infrastructure

Storage

- Local Storage

Computing

- Local systems
- AWS through AMCE

POC: svetla.m.hristova-veleva@nasa.gov
https://tropicalcyclone.jpl.nasa.gov/

Earth Data Analytics System (EDAS) as an Analytic Center

Climate Impact

- Project Definition
- Plan for Investigation

Data

- MERRA
- ECMWF ERA
- NOAA NCEP CFSR
- JMA JRA
- UA ERA

Project Work Environment -NCCS-

Tools

- Earth System Grid Federation Framework
- Discovery & Catalog
- Work Management – Jupyter Notebook
- Data Interfaces - OpenDAP
- Analytic Tools: Min, max, sum, diff, average, rms,anomaly, std deviation
- Visualization using time series or spaceplot plot routines

Computational Infrastructure

Storage

- Local Sparc
- SGI Data Migration Facility

Computing

- Local systems
- High End Computing (DISCOVER)
- GSFC Science Cloud

POC: laura.carriere@nasa.gov
https://www.nccs.nasa.gov/services/Analytics

NASA Earth Exchange (NEX) as an Analytic Center

Land Change/Use Community

- Project Definition
- Plan for Investigation

Data

- Built into website
- Dataset Sources
  - Landsat
  - Sentinel 1A
  - Modis
  - ASTER
  - TRMM
  - AVHRR
- Climate Datasets
- Landcover
- Digital Elevation Map
- STATSGO Soils
- USDA Aerial NAIP
- And others

Project Work Environment -NEX-

Tools

- Models
  - Tops
  - Biome-BGC
  - LPJ Dynamic Global Model
- Sandbox for small scale experiments
- Analysis
  - R and python based tools
  - Matlab VIIRS HDF5 swath conversion
- Workflow: Jupyter Notebook

Computational Infrastructure

Storage

- Data Containers
- Thematic model
- Metadata/Ontology
- Resulting Products
- Published data
- Provenance

Computing

- NASA Advanced Supercomputing (NAS)
- Amazon Web Services (AWS) Public Cloud
- Ames Quantum Computer (D-wave)

POC: rama.nemani@nasa.gov
https://nex.nasa.gov/nex

Recent AIST Development Projects

- Data Wrangling
  - Data Container Studies reported at AGU2015, AGU2016
  - Ontology-based metadata (AIST14-0095)
  - Science Enterprise Service Bus (AIST14-0014)
  - In Situ-Satellite matchup service (AIST14-0086)
  - SciSpark (AIST14-0034)
  - Quantum Computing tools (AIST14-0096, AIST16-0091)
  - Multi-instrument data integration (AIST11-0077, AIST16-0048)
- Workflow tools (AIST16-0092)
- Analytic Tools
  - Regression analysis on Global Vegetation (AIST14-0115)
  - Pattern-based content analysis in GIS (AIST14-0027)
  - Uncertainty Quantification (AIST11-0011, AIST16-0030)
  - Computer-aided Discovery (AIST14-0036, AIST16-0048)
- Discipline Specific Tools
  - EcoSis Tools (AIST16-0118)
  - Hyperspectral image analysis tools (AIST16-0138)
- Modeling
  - Community re-organization biodiversity models (AIST16-0052)

Needs for Data Support Tools

- Interact with data from many sources
  - Identify known data sources
  - Determine data format characteristics
  - Add local, custom or authorized restricted data sources
  - Create interoperability interfaces
  - Stabilize data so it is persistent and available for others to evaluate
  - Refine existing metadata and add to it
    - Provenance
    - Reflect additional processing
    - Precision Ontology (Beth Huffer, AIST14-0095)

Needs for Analysis Tools

- Assume data is in a tractable format
- Perform what-if exercises on data analysis
- Track the analytic process performed
  - So it can be repeated and refined
  - Helps to accurately report results in publications
- Determine relevant past work and who did it
  - Published papers
  - Abstracts of research papers at AGU, AWS, ESIP, NTIS,
  - Abstracts of research awards in US Government

Needs for the Project Work Environment

- Human Factors Engineering
- A way to view the investigation that is comfortable to PI and the team
  - How do THEY think about the problem?
  - They must believe the AC reflects their view of the problem and will produce credible and defensible results
- Support seamless orchestration of tools operating on the data
  - It must feel natural in the context of the investigation
  - Accommodate What-if exercises at each stage in using different implementations of same algorithm or different algorithms for same function
- Permit PI to plan the resources needed for the Investigation
  - Assess the cost/performance tradeoff in planning
  - Unified Access (single sign-on)
  - Monitor consumption against the plan
- It must provide an automatic record of the work
  - Repeatability and credibility
  - Accelerate accurate details in publications

Needs for the Computational Resources

- Sufficient capacity
- Pay as you
- Special capabilities
- Affordability within the project budget
  - Advance prediction of the funding needs
  - Continuous update of plan vs. actual
- Security to preclude early release
- Different stages of the investigation invariably require different resources

Modeling in the Analytic Center

- Models capture and assess knowledge and understanding of the phenomena and processes
- Assess alternative models performance
- Propagate error estimates throughout

Analytic Center in a Box

- U-haul for a science investigation
- Create an AC for an Investigation
  - Low cost
  - Low latency
  - Define the expected needs for an investigation
  - Examine sensitivity of cost to investigation parameters
  - Provide standard tool selection
  - Permit easy incorporation of custom/unique tools

Analytic Center ConOps-Initialization

- Initialization
  1. Select environment framework and notebook
  2. Outline investigation strategy in Notebook
     a. Identify, authorize and authenticate team
  3. Identify computing environment
  4. Select known data sources and interfaces
  5. Select known analytic tools
  6. Identify custom/local data sources
     a. define interface
  7. Import custom or unique tools
     a. Interface them into system
  8. Validation harmonization of all tools and data

Analytic Center ConOps-Analysis

- Seamless interaction with the data
  - Plug and play
- Tracking step-wise the analytic process
- TBD features
  - Analytic Center in a Box
    - Speed to set up
    - Ease of customizing tool entries
    - Ease of initializing data interfaces
    - Ease of initializing computing environment

Analytic Center ConOps-Publish

- Providing access to the data used upon which to base conclusions has become a necessity for publications to be accepted
- New NASA requirements for publications make distribution to all the mandatory delivery points time consuming
- AGU and the majority of the publishing industry are working on a template for submission of articles for peer review
  - Extensive supporting documentation
- The publication tool set should enable meeting these new and old bureaucratic requirements
  - Collect artifacts as the Investigation proceeds so they can be assimilated into the publication submission
  - Crediting references and data sources
  - Enabling the access to the data for validation
  - Distribution of the documents to the required distribution points
  - Configuration management
  - Reviewer comment resolution and tracking

AIST MCS End State – PI Perspective

- Relationship with AMCE is defined by a Service Level Agreement
  - PI will work with AMCE Cloud Project Manager and Cloud System Administrator to estimate their consumption of cloud resources
  - Agreement to be monitored for computer security and accounting purposes
- Each PI has access to the full range of AWS Cloud Services
  - Choices of what type and how much are left to the PI and his team
    - As new services are created by AWS, they are made available to PI as soon as possible
  - PI and his team are familiar with the services they need and how to use them in the AIST MCS
- AIST Cloud Orientation Class for PI and entire Team
  - Outline rules of engagement: consumption and usage, fiscal restraints and computer security measures
  - Emphasis on value of relationship, as defined by an agreement among all parties, not on regulatory compliance
- Reports are pushed to PI describing consumption, and logins
  - PI's are expected to review these and work with AMCE Cloud PM/SA to resolve any issues
- Financial Management
  - During early phase startup, AIST PM funds AWS and allocates funding to each Project
  - Once orientation/debugging is complete, the PI pays for computing services as he uses them, using money from his project
- AIST MCS supports both NASA Govt and Contractor employees, non-US citizens, University Researchers and commercial companies
    - PI and a lead System Administrator are Privileged Users
    - Researchers are Users

ESTO Earth Science Technology Office

Cloud Service Roles Defined

- **Billing Manager** is the cloud services team member responsible for the creation of as well as financial monitoring and oversight of all AWS linked billing accounts. (SGT AWS Account, eventually SEWP Master)
- **Master Cloud Administrator** is the cloud services team member responsible for Identity and Access Management (IAM) and technical administration and support for the AWS linked billing accounts. (SGT AWS Account, eventually ARC SEWP manager responsibility)
- **Account Coordinator** is the person within the organization responsible for requesting/establishing, funding, and first line oversight of an IAM account. (AIST Funder Role)
- **CyberSecurity Administrator** is the person responsible for the Identity and Access Management within a cloud account, establishing security policies, and establishing a separate cloud instance as a log server to log, monitor, and review all cloud instance(s) activity under their cloud account in compliance with LaRC IT Security requirements. This role may be fulfilled by either a cloud services team member or a person designated by the Account Coordinator. (AIST Role possibly combined with Cloud SA person, but separate role)
- **Cloud Systems Administrator** (SA) is the person responsible for creating cloud instances, technical administration and support, specified tagging and establishing automated billing alerts, and performing emergency stops of an instance of the authorized AWS IAM accounts. This role may be fulfilled by either cloud services team member or a person designated by the Account Coordinator. (AIST Role)
- **Cloud Project Manager** is the person within the organization responsible for providing funding to the Account Coordinator for the project and monitoring the project's usage. This is an optional role that may be designated by the Account Coordinator. (AIST Role)
- **Privileged User** is the person within the organization responsible for starting/stopping and provisioning cloud instances. This role may be fulfilled by either cloud services team member or a person designated by the Account Coordinator. (PI System Admin)
- **End User** means any individual or entity that directly or indirectly accesses or uses a cloud instance and/or cloud services under Langley Services. This person has the ability to stop cloud instances. (Project Team)
- **Technical Monitor** is the Technical Monitor (TM) for the Contractor Task Order providing cloud services. (Government procurement activity, initial LaRC Fran Risinger, then AIST, then SEWP)